

A Preliminary Study on the Learning Informativeness of Data Subsets

Simon Kaltenbacher¹, Nicholas H. Kirk² and Dongheui Lee²

Estimating the internal state of a robotic system is complex: this is performed from multiple heterogeneous sensor inputs and knowledge sources. Discretization of such inputs is done to capture saliences, represented as *symbolic* information, which often presents structure and recurrence. As these sequences are used to reason over complex scenarios [1], a more compact representation would aid exactness of technical cognitive reasoning capabilities, which are today constrained by computational complexity issues and fallback to representational heuristics or human intervention [1], [2]. Such problems need to be addressed to ensure timely and meaningful human-robot interaction.

Our work is towards understanding the variability of learning informativeness when training on subsets of a given input dataset. This is in view of reducing the training size while retaining the majority of the symbolic learning potential. We prove the concept on human-written texts, and conjecture this work will reduce training data size of sequential instructions, while preserving semantic relations, when gathering information from large remote sources [3].

Posterior Evaluation Distribution of Subsets

We computed multiple random subsets of sentences from the UMBC WEBBASE CORPUS (~ 17.13GB) via a custom implementation using the SPARK distributed framework. We evaluated the learning informativeness of such sets in terms of semantic word-sense classification accuracy (with WORD2VEC [4]), and of n-gram perplexity. Previous literature inform us that corpus size and posterior quality do not follow linear correlation for some learning tasks (e.g. semantic measures) [5]. In our semantic tests, on average 85% of the quality can be obtained by training on a random ~ 4% subset of the original corpus (e.g. as in Fig. 1, 5 random million lines yield 64.14% instead of 75.14%).

Our claims are that i) such evaluation posteriors are Normally distributed (Tab. I), and that ii) the variance is inversely proportional to the subset size (Tab. II).

It is therefore possible to select the best random subset for a given size, if an information criterion is known. Such metric is currently under investigation. Within the robotics domain, in order to reduce computational complexity of the training phase, cardinality reduction of human-written instructions is particularly important for non-recursive online training algorithms, such as current symbol-based probabilistic reasoning systems [1], [3], [6].

¹S.K. is with Ludwig Maximilian University of Munich, Germany simon.kaltenbacher@campus.lmu.de

²N.H.K. and D.L. are with the Technical University of Munich, Germany {nicholas.kirk,dhlee}@tum.de

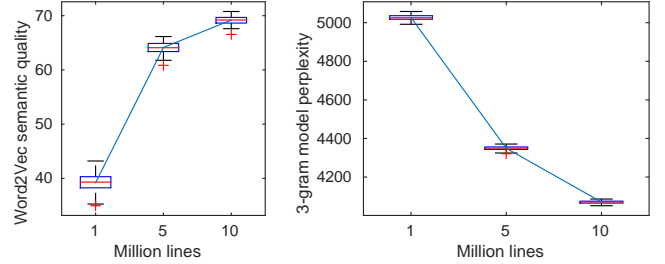


Fig. 1. Evaluation values for random subsections of various sizes, for both semantic and syntactic tasks (100 instances for each visualized size).

		100 subsets of 1M		100 subsets of 5M		100 subsets of 10M	
		h	p	h	p	h	p
WORD2VEC	χ^2	0	0.4221	0	0.5756	0	0.9189
	And.-Darling	0	0.8749	0	0.7616	0	0.8710
PERPLEXITY	χ^2	0	0.2963	0	0.2435	0	0.2443
	And.-Darling	0	0.4908	0	0.1488	0	0.3423

TABLE I

CHI-SQUARE AND ANDERSON-DARLING TESTS SHOWING THERE IS NO GAUSSIAN NULL HYPOTHESIS REJECTION FOR WORD2VEC AND PERPLEXITY ACCURACY VALUES OF RANDOM SUBSETS (10% SIGNIFICANCE LEVEL).

		100 subsets of 1M variance	100 subsets of 5M variance	100 subsets of 10M variance
WORD2VEC		2.6199	1.0351	0.6147
PERPLEXITY		213.21	118.87	55.218

TABLE II

VARIANCE VALUES OF WORD2VEC AND PERPLEXITY ACCURACY POSTERiors OF RANDOM SUBSETS.

REFERENCES

- [1] N. H. Kirk, K. Ramirez-Amaro, E. Dean-Leon, M. Saveriano, and G. Cheng, "Online prediction of activities with structure: Exploiting contextual associations and sequences," in *2015 IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2015.
- [2] N. H. Kirk, D. Nyga, and M. Beetz, "Controlled natural languages for language generation in artificial cognition," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6667–6672, IEEE, 2014.
- [3] M. Tenorth, D. Nyga, and M. Beetz, "Understanding and executing instructions for everyday manipulation tasks from the world wide web," in *2010 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1486–1491, IEEE, 2010.
- [4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, pp. 3111–3119, 2013.
- [5] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pp. 26–33, Association for Computational Linguistics, 2001.
- [6] N. H. Kirk, "Towards learning object affordance priors from technical texts," in *"Active Learning in Robotics" Workshop, 2014 IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2014.